GTN-Québec
Groupe de travail québécois sur les normes et standards
en TI pour l'apprentissage, l'éducation et la formation

Publication 2011-07

Évaluation des métadonnées extraites par ExifTool aux fins de création d'une fiche LOM

Marc-Antoine Parent

Mission du GTN-Québec

La mission du Groupe de travail québécois sur les normes et standards pour l'apprentissage, l'éducation et la formation (GTN-Québec) est de fournir une expertise à la communauté éducative en matière de normalisation.

Les membres du GTN-Québec proviennent des trois ordres d'enseignement, des ministères, ainsi que du secteur privé de la formation. En s'appuyant sur les travaux des groupes internationaux d'élaboration des normes, ils soutiennent les acteurs du milieu de l'éducation pour favoriser l'implantation de pratiques communes de description et de production de ressources éducatives interopérables, réutilisables et accessibles à tous.

Ces ressources forment un patrimoine éducatif d'une valeur inestimable pour les communautés éducatives francophones. Assurer son enrichissement et sa pérennité est en conséquence, depuis sa fondation, au cœur des préoccupations du GTN-Québec.

Objectifs du GTN-Québec

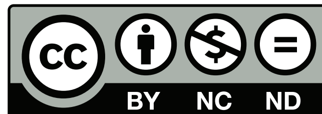
1. Dans une perspective d'accompagnement, consulter les acteurs du milieu de l'éducation pour mieux définir comment les approches basées sur les normes et standards peuvent aider à concrétiser la mission éducative de leur organisation ;
2. Connaître des solutions basées sur des normes et standards, s'assurer qu'elles correspondent à la réalité et aux besoins du milieu et proposer, le cas échéant, des adaptations ou des guides d'utilisation de ces normes ;
3. Faire connaître et encourager les pratiques normalisées de production et de description de ressources éducatives ;
4. Favoriser le développement d'une masse critique de REA numériques accessibles, pérennes et réutilisables au sein des établissements de chaque ordre d'enseignement ;
5. Maintenir l'expertise et la représentation québécoises en matière de développement de normes internationales et d'autres standards.

Les activités du GTN-Québec sont réalisées avec l'appui financier du ministère de l'Éducation, du Loisir et du Sport du Québec et grâce à la collaboration de ses membres.

www.gtn-quebec.org

ISBN 978-2-924168-16-5
Dépôt légal – Bibliothèque et Archives nationales du Québec, 2011
Dépôt légal – Bibliothèque et Archives Canada, 2011

Licence de la propriété intellectuelle



Cette étude a été réalisée avec le soutien financier du Groupe de travail québécois sur les normes et standards en TI pour l'apprentissage, l'éducation et la formation (GTN-Québec). Le contenu de ce rapport demeure la responsabilité des auteurs. Les opinions qui y sont exprimées ne reflètent pas nécessairement celles du GTN-Québec.

Auteur

Marc-Antoine Parent, M.Sc., est responsable du secrétariat technologique du GTN-Québec. Il a travaillé dans l'industrie des T.I. et y a souvent occupé des fonctions de recherche. Il assure la veille technologique et le transfert au GTN-Québec, notemment sur les standards de métadonnées.

Table des matières

Licence de la propriété intellectuelle 1

Auteur 2

Table des matières 3

Objectif : capture des métadonnées contenues dans les documents courants 4

Méthodologie 5
- Base de tests, métadonnées présentes 5
- Détection et résolution de conflits 5
- Outil de génération de LOM 5

Travaux futurs 7

Appendice : Tableau-synthèse. 8

Objectif : capture des métadonnées contenues dans les documents courants

Ces travaux font suite au travail de débroussaillage réalisé par Vincent François, de WebConforme, pour identifier les métadonnées présentes dans les documents, et pouvant être extraites par Alfresco. Nous avions relevé qu'Alfresco sélectionnait une partie des métadonnées détectées par l'outil ExifTool [1] de Phil Harvey, et qu'il serait pertinent de voir si un emploi direct d'ExifTool nous permettrait de récupérer plus de métadonnées, et d'en tirer une fiche LOM. Nous nous sommes donc concentrés sur les métadonnées contenues dans LOM.

Nous avons également relevé qu'ExifTool permettait la manipulation des métadonnées dans certains cas, et envisagé de nous en servir pour enrichir les documents. En effet, ExifTool permet, dans certains cas, d'éditer les valeurs de métadonnées contenues dans les documents. Toutefois, nous avons dû abandonner ce second objectif, car l'édition est surtout gérée pour les formats graphiques, grâce à la norme Exif [2], mais ce type de fichiers n'est pas des plus importants pour les REA. Il est également possible d'éditer des métadonnées génériques XMP [3], telles que présentes dans les types de documents définis par Adobe, y compris PDF. En théorie, des métadonnées XMP peuvent être intégrées dans toutes sortes de formats binaires ; mais en pratique, ExifTool ne se risque pas à ajouter des données XMP à des formats binaires qui ne sont pas prévus à cet effet. Donc, ExifTool ne permet pas d'édition de données dans les principaux formats de documents éditables, soit les suites Microsoft Office, OpenOffice ; ni d'éditer les métadonnées dans un document HTML. On peut consulter la liste des formats lus et écrits [4] dans la documentation d'ExifTool.

1. http ://www.sno.phy.queensu.ca/∼phil/exiftool/
2. http ://exif.org/
3. http ://www.adobe.com/products/xmp/
4. http ://www.sno.phy.queensu.ca/∼phil/exiftool/#supported

Méthodologie

Base de tests, métadonnées présentes

Nous sommmes partis, pour la base de tests, des documents de tests fournis avec ExifTool. Nous y avons adjoints quelques fichiers *ad hoc*, en particulier des documents Microsoft Office, Open Office et iWorks, qui ne sont pas tous dans la suite de texte fournie, ou qui ne contenaient pas autant de métadonnées que possible. Cette liste ne correspond pas exactement aux fichiers employés dans l'étude de WebConforme, mais contient les principaux types mentionnés et pourra être enrichie avec des exemples. Nous avons appliqué ExifTool à chacun de ces fichiers pour en extraire les métadonnées, et avons sélectionné celles qui pouvaient correspondre à des champs LOM, en regardant à la fois le nom de la métadonnée et son contenu dans les fichiers de test.

Détection et résolution de conflits

Dans de nombreux cas, ExifTool retrouve plusieurs métadonnées pouvant servir de base à un même champ LOM, parfois même dans un même fichier. (Par exemple, pas moins de 30 métadonnées différentes pour la date, et une dizaine d'autres dont le nom contient `Date` mais qui ne sont pas pertinents, comme par exemple `PatientBirthDate`) Nous avons donc automatisé l'identification des fichiers où plusieurs valeurs venant de métadonnées différentes entraient en conflit pour une seule métadonnée LOM dans un fichier donné, et avons tenté d'évaluer quelle(s) métadonnée(s) devaient prévaloir. Dans certains cas, plusieurs valeurs sont présentées et rien n'indique la valeur correcte, et nous avons dû prendre une décision plausible. Nous avons donc tenté de regrouper les métadonnées en ordre de priorité : par exemple, une donnée de type `Title` a prévalence sur le `Filename`.

Notons que certains concepts n'ont pas d'équivalent évident dans les métadonnées fournies ; par exemple, le rôle de Contributeur pourrait correspondre aux utilisateurs du mécanisme de révision dans Microsoft Office, mais seul le dernier éditeur est relevé par ExifTool.

Outil de génération de LOM

À partir de la liste ordonnée des métadonnées, nous pouvons générer un fichier XSLT permettant la transformation automatisée des données fournies par ExifTool (sous forme de XML) en une fiche LOM partiellement remplie, contenant les métadonnées extraites du document. Il serait souhaitable de pouvoir dire que ce fichier XSLT est suffisant pour opérer la conversion ; toutefois, certains cas limite nous ont obligé à employer des extensions à XSLT (entre autres, ExifTool code parfois de l'information binaire en format Base64.) Nous avons donc une implantation hybride basée sur XSLT et perl. La première version

de l'outil, encore employée pour la base de tests, utilise également python et lxml ; mais nous avons jugé préférable de recoder les composantes essentielles en perl, comme ExifTool lui-même, pour faciliter la distribution.

Cet outil est disponible comme logiciel libre sur GitHub [5]. L'outil est utilisable en commande ligne, et est capable de traiter des fichiers ou des URLs. L'outil a également été emballé sous d'application pour MacOSX 10.7+ à l'aide du logiciel Platypus [6]. Cette version emballée est disponible ici [7]. La création d'une version windows pourrait être envisagée si le besoin s'en fait sentir.

5. https ://github.com/GTN-Quebec/exif2lom
6. http ://sveinbjorn.org/platypus
7. https ://github.com/downloads/GTN-Quebec/exif2lom/Exif2LOM.dmg

Travaux futurs

Il est encore souhaitable d'étendre la base de test, mais cela peut être fait au fur et à mesure. Il serait peut-être souhaitable d'étendre les mécanismes de tests pour inclure les valeurs attendus pour certains fichiers de tests. À court terme, il serait souhaitable de rendre l'outil disponible à un public plus large ; cela pourrait impliquer d'ouvrir le source, de mieux le documenter, etc. ou à tout le moins d'en faire une version sous forme d'exécutable simple, qui générerait le LOM en y faisant glisser un fichier quelconque.

À plus long terme, il serait intéressant d'explorer des variantes basées sur une génération d'autres formats de métadonnées, tels MLR. Même si ExifTool est très limité en ce sens, il demeurerait pertinent de garder un oeil ouvert sur les outils d'édition de métadonnées. L'idéal serait de développer un outil qui assisterait les utilisateurs à voir, corriger et même enrichir les métadonnées enchâssées dans leurs fichiers, au moins lorsque c'est possible. Ainsi, nous pourrions encourager les producteurs de documents à définir des métadonnées correctes dans les documents.

En effet, il ne faut pas se leurrer : les métadonnées contenues dans les documents sont souvent invalides. Dans certains cas, il pourrait être intéressant de faire des études pour évaluer cette qualité, en comparant par exemple la langue donnée en métadonnée avec la langue reconnue dans un texte. Mais c'est là un tout autre projet.

Appendice : Tableau-synthèse.

Voici la liste des métadonnées extraites à partir de nos fichiers de test. Certaines des métadonnées étaient absentes des résulats antérieurs, notamment les dates, la taille du fichier, les informations relatives au lieu, et le texte de copyright. Quant aux autres, nous pouvons parfois récupérer des données non-reconnues dans l'étude précédente, probabablement parce qu'Alfresco ignorait quelle méta-donnée tirer de ExifTool. (Par exemple, nous reconnaissons la langue plus souvent qu'Alfresco.)

Table 1: Données identifiées par fichier de test

Fichier	title	language	description	keyword	create_date	modify_date	metadata_date	FN_lastauthor	FN_author	FN_contributor	COUNTRY	CITY	STATE	ORG	size	format	duration	copydescription	Total
Olympus.jpg	x				x										x	x			4
InDesign.indd	x				x	x									x	x			5
PSP.psp	x		x		x	x									x	x		x	7
Font.afm	x				x	x									x	x			5
FLAC.flac	x					x									x	x		x	5
LNK.lnk	x		x		x	x									x	x			6
Geotag.xml	x					x									x	x			4
DICOM.dcm	x				x	x									x	x			5
MXF.mxf	x				x	x									x	x	x		6
Matroska.mkv	x	x			x	x									x	x	x		7
Casio.jpg	x				x										x	x			4
GPS.jpg	x			x	x	x			x		x	x	x		x	x		x	11
Nikon.jpg	x				x										x	x			4
M2TS.mts	x				x	x									x	x			5
Ricoh.jpg	x				x										x	x		x	5
ZIP.zip	x					x									x	x			4
Panasonic.rw2	x				x										x	x			4
BigTIFF.btf	x					x									x	x			4
XMP.jpg	x		x		x	x		x	x		x	x	x		x	x		x	12
ExtendedXMP.jpg	x				x			x	x						x	x			6
Sanyo.jpg	x				x										x	x			4
TestWord11.pdf	x				x			x	x						x	x			6
Sigma.jpg	x				x	x									x	x			5
MIE.mie	x			x	x	x					x	x	x		x	x		x	10
.DS_Store	x					x									x				3
GE.jpg	x				x	x									x	x			5
Sony.jpg	x				x										x	x			4
XMP2.xmp	x		x			x									x	x			5
Real.ram	x					x									x	x			4

suite page suivante

Fichier	title	language	description	keyword	create_date	modify_date	metadata_date	FN_lastauthor	FN_author	FN_contributor	COUNTRY	CITY	STATE	ORG	size	format	duration	copydescription	Total
FujiFilm.raf	x				x										x	x		x	5
ZIP.gz	x					x									x	x			4
AIFF.aif	x				x	x			x						x	x			6
Geotag.log	x					x									x				3
Olympus2.jpg	x				x										x	x			4
Alfresco-Normetic Rapport final–1.pdf	x	x			x	x		x	x						x	x			8
JVC.jpg	x				x										x	x			4
Sigma.x3f	x				x	x									x	x			5
CanonRaw.cr2	x				x										x	x			4
MWG.jpg	x		x	x	x	x			x		x	x	x		x	x		x	12
iWork.numbers	x			x		x			x						x	x		x	7
Font.ttf	x					x									x	x		x	5
PICT.pict	x					x									x	x			4
EXE.macho	x					x									x	x			4
Canon.jpg	x				x										x	x			4
ExifTool.tif	x			x	x	x			x		x	x	x		x	x		x	11
GeoTiff.tif	x					x									x	x			4
CasioQVCI.jpg	x				x	x									x	x			5
Nikon.nef	x				x	x					x	x	x		x	x			8
Font.pfb	x				x	x			x						x	x			6
Font.pfa	x				x	x			x						x	x			6
TestWord11_ed.docx	x	x	x	x	x	x		x	x					x	x	x			11
Vorbis.ogg	x				x	x									x	x			5
EXE.exe	x	x	x			x									x	x		x	7
Font.pfm	x					x									x	x		x	5
PGF.pgf	x					x									x	x			4
Kodak.jpg	x				x	x									x	x			5
DjVu.djvu	x		x	x	x	x		x	x						x	x		x	10
Minolta.mrw	x				x										x	x			4
AFCP.jpg	x			x	x	x			x		x	x	x		x	x		x	11
Real.ra	x					x									x	x		x	5

suite page suivante

Fichier	title	language	description	keyword	create_date	modify_date	metadata_date	FN_lastauthor	FN_author	FN_contributor	COUNTRY	CITY	STATE	ORG	size	format	duration	copydescription	Total
FotoStation.jpg	x					x									x	x			4
KyoceraRaw.raw	x				x	x									x	x			5
FlashPix.ppt	x		x	x	x	x		x	x					x	x	x			10
IPTC.jpg	x			x	x	x			x		x	x	x		x	x		x	11
Casio2.jpg	x				x							x			x	x			5
DV.dv	x				x	x									x	x	x		6
FujiFilm.jpg	x				x										x	x		x	5
Jpeg2000.j2c	x					x									x	x			4
PDF.pdf	x		x	x	x	x		x	x		x	x	x		x	x		x	13
TestWord11.doc	x	x	x	x	x			x	x					x	x	x			10
TestWord11.htm	x	x	x	x	x			x	x					x	x	x			10
Sony.pmp	x				x										x	x			4
APE.ape	x					x									x	x			4
Flash.swf	x					x			x						x	x	x		6
Real.rm	x		x	x	x	x			x						x	x	x	x	10
CanonRaw.crw	x				x	x									x	x			5
test_openoffice_texte.odt	x		x		x	x			x						x	x	x		8
CaptureOne.eip	x				x	x									x	x			5
APE.mpc	x				x	x									x	x			5
JVC2.jpg	x				x										x	x			4
Unknown.jpg	x				x										x	x			4
ITC.itc	x					x									x	x			4
Pentax.avi	x				x	x									x	x	x	x	7
EXE.elf	x					x									x	x			4
CanonVRD.vrd	x					x									x	x			4
XMP.xml	x				x										x	x			4
HTML.html	x	x	x	x	x	x		x	x	x				x	x	x	x		13
Minolta.jpg	x				x										x	x			4
OOXML.docx	x	x	x	x	x	x		x	x					x	x	x			11
XMP3.xmp	x					x			x		x				x	x			6
NikonD2Hs.jpg	x				x										x	x			4
Flash.flv	x					x	x								x	x	x		6

suite page suivante

Fichier	title	language	description	keyword	create_date	modify_date	metadata_date	FN_lastauthor	FN_author	FN_contributor	COUNTRY	CITY	STATE	ORG	size	format	duration	copydescription	Total
Jpeg2000.jp2	x					x									x	x			4
PDF2.pdf	x					x									x	x			4
ASF.wmv	x				x	x								x	x	x	x		7
PostScript.eps	x		x	x	x	x		x	x		x	x	x		x	x		x	13
MP3.mp3	x				x	x									x	x	x		6
GIF.gif	x					x									x	x			4
BMP.bmp	x					x									x	x			4
PhotoMechanic.jpg	x		x	x	x	x		x	x		x	x	x		x	x		x	13
Photoshop.psd	x		x			x		x	x						x	x		x	8
XMP5.xmp	x			x		x									x	x			5
RTF.rtf	x		x	x		x			x					x	x	x		x	9
RIFF.avi	x				x	x	x		x						x	x	x		8
Geotag.gpx	x					x									x	x			4
TestWord11.rtf	x	x	x	x	x			x	x					x	x	x			10
MIFF.miff	x		x	x	x	x		x	x		x	x	x		x	x		x	13
FLAC.ogg	x					x									x	x			4
RIFF.wav	x				x	x									x	x	x		6
TestWord11.docx	x	x	x	x	x	x		x	x					x	x	x			11
PNG.png	x					x			x						x	x			5
GIMP.xcf	x		x			x			x						x	x		x	7
QuickTime.m4a	x				x	x									x	x	x		6
Writer.jpg	x					x									x	x			4
DNG.dng	x				x	x			x						x	x			6
Pentax.jpg	x				x							x			x	x			5
PPM.ppm	x					x									x	x			4
NikonD70.jpg	x				x										x	x			4
ExifTool.jpg	x		x	x	x	x		x	x		x	x	x		x	x		x	13
Panasonic.jpg	x				x										x	x			4
XMP.svg	x	x	x		x	x			x					x	x	x			9
SigmaDP2.x3f	x				x										x	x			4
Geotag.igc	x					x									x				3
OlympusE1.jpg	x				x										x	x			4

suite page suivante

Fichier	title	language	description	keyword	create_date	modify_date	metadata_date	FN_lastauthor	FN_author	FN_contributor	COUNTRY	CITY	STATE	ORG	size	format	duration	copydescription	Total
QuickTime.mov	x				x	x	x		x						x	x	x		8
XMP.xmp	x		x		x	x	x								x	x		x	8
Font.dfont	x					x									x	x		x	5
Ricoh2.jpg	x				x										x	x		x	5
Canon1DmkIII.jpg	x				x										x	x			4
XMP4.xmp	x		x			x				x	x	x			x	x			8
OpenDoc.ods	x		x		x	x			x						x	x	x		8